Computing and Artificial Intelligence 2025, 3(4), 3893.
doi: https://doi.org/10.59400/cai3893

Article

Verifying artificial intelligence-generated images: Socio-technical approaches to authenticity

Michael Mncedisi Willie

Policy Research and Monitoring, Council for Medical Schemes, Pretoria 0157, South Africa; m.willie@medicalschemes.co.za

CITATION

Willie MM. Verifying artificial intelligence-generated images: Socio-technical approaches to authenticity. Computing and Artificial Intelligence. 2025; 3(4): 3893. https://doi.org/10.59400/cai3893

ARTICLE INFO

Received: 25 July 2025
Revised: 30 August 2025
Accepted: 2 September 2025
Available online: 5 October 2025

Abstract: The rapid proliferation of artificial intelligence (AI) has transformed visual media, enabled highly realistic AI-generated images, and raised ethical, social, and security concerns. Generative artificial intelligence (Generative AI) architectures, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, allow content creation that is increasingly indistinguishable from human-made visuals, facilitating creativity, education, and communication. However, these capabilities also introduce risks of manipulation, identity fraud, misinformation, and deepfake attacks across social, political, corporate, academic, and humanitarian domains. This study investigates AI image verification as a socio-technical response to synthetic visuals, focusing on social media, artistic, and forensic contexts. It employed a qualitative design combining thematic literature review and case study analysis. Thematic analysis identified patterns in verification approaches, including pixel-level analysis, metadata forensics, machine learning classifiers, watermarking, and blockchain-enabled methods. Case studies explored real-world applications, highlighting perceptual biases, strategic use of synthetic content, and governance and digital literacy challenges. Findings reveal that human perception alone is insufficient for reliably discerning authenticity, with individuals frequently misclassifying AI-generated images as real. Integrating machine learning, metadata analysis, and blockchain verification, hybrid technical approaches significantly improve detection accuracy. Socio-technical factors, including platform policies, ethical norms, organisational governance, and user literacy, shape the effectiveness of verification methods. The study presents a conceptual framework linking technological, organisational, and societal dimensions, emphasising the need for coordinated strategies that combine algorithmic innovation, regulatory oversight, and public engagement. Practical implications include deploying hybrid verification systems, strengthening governance and ethical standards, enhancing digital literacy, and fostering cross-disciplinary collaboration to safeguard trust, authenticity, and integrity in digital media.

Keywords: AI-generated images; deepfake detection; social media manipulation; forensic analysis; hybrid verification; digital literacy; socio-technical systems; image authenticity

1. Introduction

The rapid proliferation of artificial intelligence (AI) has profoundly transformed the production and dissemination of visual media, introducing unprecedented efficiencies, personalisation, and precision in content creation [1,2]. These developments highlight AI's potential to drive innovation across creative and industrial domains and foreground critical ethical and authenticity challenges that demand scrutiny [1]. AI-generated visuals

redefine digital creativity by combining human imagination with machine precision, enabling expansive experimentation, and reshaping conventional notions of artistic expression [3,4].

Generative AI architectures, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models, are producing images of striking realism, increasingly indistinguishable from genuine photography [5]. Integrating complementary technologies, such as speech-to-text translation with GAN-based systems, expands creative and educational applications by generating diverse, high-fidelity visuals from spoken prompts [6]. While these capabilities enrich communication and creative output, they simultaneously heighten the risk of misuse for manipulation, identity theft, and societal misinformation, necessitating robust verification mechanisms spanning image forensics, metadata analysis, and deepfake detection [7,8].

The societal implications of generative AI extend beyond technical challenges, raising pressing ethical concerns regarding privacy, originality, and the potential marginalisation of human creative labour [9, 10]. In domains where authenticity is paramount, such as journalism, politics, finance, and security, the capacity to fabricate convincing visuals underscores the urgent need for reliable detection frameworks. Recent advances in forensic AI, including deep learning-based classifiers and metadata analysis, have demonstrated substantial accuracy in identifying synthetic content, offering practical tools for safeguarding privacy and verifying media integrity [11–13]. This study examines AI image verification as a technical and strategic response to synthetic visuals. It evaluates detection methods, including pixel analysis, metadata forensics, and machine learning classifiers, assessing their effectiveness and limitations.

2. Literature review

2.1. Generative visual AI

Generative visual AI has evolved rapidly, from basic image synthesis to creating hyper-realistic visuals that closely emulate real-world scenes. Leveraging architectures such as GANs, VAEs, and diffusion models, generative AI is reshaping sectors ranging from computer vision and healthcare to creative industries, while simultaneously raising ethical, security, and deepfake-related concerns [14,15]. Tools such as DALL·E, Midjourney, Stable Diffusion, and LeonardoAI exploit large-scale datasets to learn and reproduce complex visual patterns, enabling rapid, contextually informed content creation that fosters innovation in design, education, and digital media [16,17].

Despite these transformative capabilities, democratising generative AI also lowers barriers for malicious use, amplifying misinformation, fraud, and deception risks. Deepfake technologies, powered by generative AI, exemplify these challenges, exposing vulnerabilities in security, trust, and verification processes, and underscoring the need for interpretable, real-time detection frameworks [18]. Moreover, while text-to-image AI tools facilitate rapid ideation and urban design visualisation, their effectiveness is context-dependent and often requires complementarity with traditional design methods [17]. Building on the ethical and security concerns of generative AI, it is important to explore how these powerful tools can also create real-world

risks, primarily through synthetic visuals that may be misused in social, political, and economic contexts.

2.2. Synthetic visual risks

The proliferation of synthetic visuals presents multifaceted risks across social, political, and economic spheres. Deepfake images and videos have been deliberately weaponised in political disinformation campaigns, manipulated to mislead voters or erode public trust, exposing significant vulnerabilities in democratic processes. Public perceptions of AI-generated deepfakes vary widely by age, country, and education, with risks often perceived as outweighing benefits, underscoring the necessity for context-specific governance and detection strategies [19]. In financial domains, synthetic visuals facilitate fraud through executive impersonation, fabricated documentation, and market manipulation, highlighting the need for integrated technological and regulatory countermeasures [20]. Individuals also face heightened threats to privacy and identity, as AI-generated synthetic identities leveraging models such as GANs and large language models can convincingly mimic real users on social media, undermining trust, and authenticity [21]. Proposed frameworks, such as the business privacy calculus, demonstrate how AI-driven systems can mitigate deepfake risks, safeguard data integrity, and manage operational vulnerabilities in sensitive sectors [22]. Romero-Moreno further argues that effectively addressing deepfake threats requires a synthesis of AI-powered detection, ethical oversight, adaptive governance, and international legal harmonisation to protect human rights while maintaining media credibility [23]. Traditional verification approaches, including human scrutiny, metadata analysis, and provenance checks, are increasingly insufficient against sophisticated outputs capable of simulating natural imperfections. This highlights a critical gap between generative innovation and the mechanisms needed to safeguard authenticity.

2.3. Detection approaches

The literature identifies various technical strategies for detecting synthetic visuals, each presenting unique strengths and limitations. Pixel-level analysis, for instance, focuses on anomalies in textures, reflections, or lighting patterns; however, its efficacy diminishes as advanced generative models increasingly replicate natural imperfections. Hybrid approaches, combining stylometric analysis, watermarking, pixel prediction, machine learning, and blockchain, have achieved detection accuracies of up to 92%, while also addressing ethical, privacy, and regulatory considerations [24].

Deepfake video detection remains particularly challenging due to unbalanced or poorly labelled datasets, high computational demands, overconfidence in detection models, and limited generalisation of deep learning methods [18,19]. These limitations underscore the necessity for robust, real-time detection mechanisms supported by high-quality, diverse datasets. Metadata forensics, such as examining EXIF data, can expose inconsistencies; however, metadata can be easily altered or stripped, limiting its reliability. Machine learning classifiers offer adaptive detection by learning subtle generative artefacts, though continuous retraining is required to remain effective against

evolving AI outputs. Digital image forensics, leveraging EXIF metadata, has proven useful in authenticating images and supporting investigative workflows [25].

Support Vector Machine (SVM) models based on Discrete Fourier Transform (DFT) analyses demonstrate high accuracy in distinguishing genuine from manipulated digital images and videos while reducing processing time relative to some deep learning approaches, making them suitable for integration into digital forensics platforms [26]. Emerging techniques, such as blockchain-based verification and digital watermarking, embed authenticity markers at content creation, offering resilience against manipulation but raising concerns around scalability and creative freedom. Integration of blockchain with watermarking and perceptual hash functions enhances content verification, reduces computational overhead, and supports tamper-resistant video copyright protection without necessitating full storage on the blockchain [27, 28]. Furthermore, blockchain-enabled watermarking improves deepfake detection through immutable, decentralised verification, though challenges persist regarding computational cost and standardisation.

While these technical strategies represent substantial progress, detection methods continue to lag behind the rapid evolution of generative AI. This gap highlights the need for continuous research, cross-disciplinary collaboration, and the integration of both technical and regulatory solutions to strengthen authenticity verification and safeguard public trust in digital media [29].

3. Theoretical framework

This study's theoretical framework is grounded in media forensics and socio-technical systems theory, providing an integrated lens to critically examine the intersection of AI-generated visual content, detection technologies, and societal consequences. Media forensics offers conceptual and methodological tools to detect, analyse, and verify digital imagery, focusing on artefacts, anomalies, and manipulations inherent in synthetic visuals. In particular, deepfake forensics, including passive and active authentication and attribution techniques, addresses emergent challenges such as misinformation, fraud, and the "Impostor Bias," thereby ensuring media integrity in practical contexts [30]. Systematic frameworks for digital forensics further enhance the collection, analysis, and management of digital evidence, addressing gaps in conventional models and supporting more robust cybercrime investigations [31].

Complementing this, socio-technical systems theory situates AI image verification within broader organisational and societal contexts, emphasising the co-evolution of technological capabilities, human expertise, social structures, and regulatory frameworks. While AI techniques such as natural language processing and graph neural networks offer promising applications in social media forensics for detecting cybercrime, misinformation, and harmful online behaviour, their effectiveness is contingent upon scalability, generalisability, and interpretability [32]. Similarly, using social media within socio-technical systems can enhance engagement and improve access to services in health, education, and smart cities. However, many implementations overlook the social dynamics and human interactions that shape these systems, which can reduce their overall effectiveness [33].

The framework further recognises the interplay between digital innovation and organisational transformation, highlighting that AI adoption, particularly in public administrations, requires systemic changes in routines, governance, culture, and structures to realise benefits while mitigating risks [34]. Moreover, integrating perspectives from mobile business model analysis underscores how interdependencies among technological, strategic, and organisational dimensions shape adoption outcomes, providing transferable insights into optimisation under dynamic, high-risk environments [35]. These theoretical lenses position AI image verification as both a technical and socio-technical problem, necessitating coordinated strategies that combine algorithmic innovation, ethical oversight, policy frameworks, and public literacy to address the risks and complexities of synthetic imagery [36].

4. Methodology

This study adopts a qualitative evidence-synthesis design, combining a structured literature review with illustrative case analysis to examine AI-generated image verification within social media and related digital contexts. The methodological approach is guided by principles of transparency, rigour, and reproducibility, and is informed by socio-technical systems theory, media forensics, and responsible AI scholarship [36]. Rather than advancing a purely technical evaluation, the study integrates technical, social, ethical, and governance dimensions to reflect the complex, real-world conditions under which AI-generated visuals are produced, disseminated, and interpreted.

To enhance methodological robustness and reporting transparency, the review process was informed by established systematic review principles, drawing on elements of PRISMA-style guidance for literature identification, screening, eligibility assessment, and inclusion. While the study does not claim a full meta-analytic systematic review, the structured procedures adopted ensure traceability, replicability, and clarity in study selection and synthesis.

Thematic analysis was employed to systematically identify, compare, and interpret recurring patterns across the literature, enabling synthesis across heterogeneous study designs and disciplinary perspectives. This approach is particularly suited to emerging and interdisciplinary fields, such as AI image verification, where empirical methods, evaluative metrics, and normative considerations vary widely. Complementing the literature synthesis, illustrative case analyses were used to ground theoretical insights in observable real-world applications, highlighting interactions among AI detection tools, user behaviour, and platform governance mechanisms [37].

Social media images were selected as the primary analytical focus due to their high velocity of dissemination, minimal gatekeeping, and heightened vulnerability to manipulation, which amplify risks of misinformation, reputational harm, and identity fraud. Prior research indicates that public perceptions of AI-generated deepfakes vary significantly across demographic, cultural, and regional contexts, reinforcing the importance of governance frameworks, regulatory oversight, and digital literacy as moderating factors in verification effectiveness [19].

4.1. Literature search and inclusion criteria

A comprehensive literature search was conducted across Scopus, Web of Science, IEEE Xplore, and Google Scholar, covering the period 2011 to 2025, to capture both foundational research and the most recent developments in generative AI and detection technologies. The studies summarized in **Table 1** show a clear evolution of research on AI-generated visual content, verification techniques, and governance frameworks over time. During 2011–2015 (n = 2, 4.9%), research was limited and focused mainly on foundational theoretical frameworks and early digital forensics concepts, published in journal articles. The period 2016–2020 (n = 2, 4.9%) remained modest in volume but introduced socio-technical and methodological perspectives, again mostly in journals.

Table 1. Summary of included studies by year range and source type.

Year range	Number of studies	Percentage (%)	Type of sources
2011–2015	2	4.9	Journal Articles
2016–2020	2	4.9	Journal Articles
2021–2025	37	90.2	Journal Articles, Conference Proceedings, Book Chapters, Technical Reports, Repository Publications
Total	**41**	**100**	

Research activity increased sharply from 2021–2025 (n = 37, 90.2%), reflecting the rapid growth of generative AI, deepfake detection, ethical considerations, and governance frameworks. Publications during this period appeared across journal articles, conference proceedings, book chapters, technical reports, and repository publications, indicating a broadening of both disciplinary engagement and publication formats. Overall, the table illustrates a trend from conceptual and methodological foundations toward applied, technical, and governance-focused studies, highlighting the field's accelerating development.

Search terms included combinations of: "AI image verification," "deepfake detection," "synthetic media," "AI-generated images," "social media misinformation," "AI governance," and "responsible AI." Boolean operators (AND, OR) and truncation strategies were applied to optimise recall and precision. The search targeted peer-reviewed journal articles, conference proceedings, policy reports, and authoritative institutional publications that addressed technical verification methods, socio-technical implications, or governance and regulatory frameworks relevant to AI-generated visual content.

4.2. Screening and selection

Figure 1 presents the PRISMA flow diagram outlining the study selection process for the qualitative evidence synthesis. A total of 101 records were initially identified through systematic database searches, including Scopus (n = 41), Web of Science (n = 32), IEEE Xplore (n = 15), and Google Scholar (n = 13). Following the removal of 1 duplicate record and 0 non-analytical or non-peer-reviewed sources, 100 unique records remained for title and abstract screening. During this phase, 59 records were excluded

for failing to meet the predefined relevance and inclusion criteria. Consequently, 41 full-text articles were retrieved and assessed for eligibility. No additional exclusions were made at this stage, and all 41 studies satisfied the methodological and thematic requirements of the review. These studies were therefore retained and included in the final qualitative synthesis.

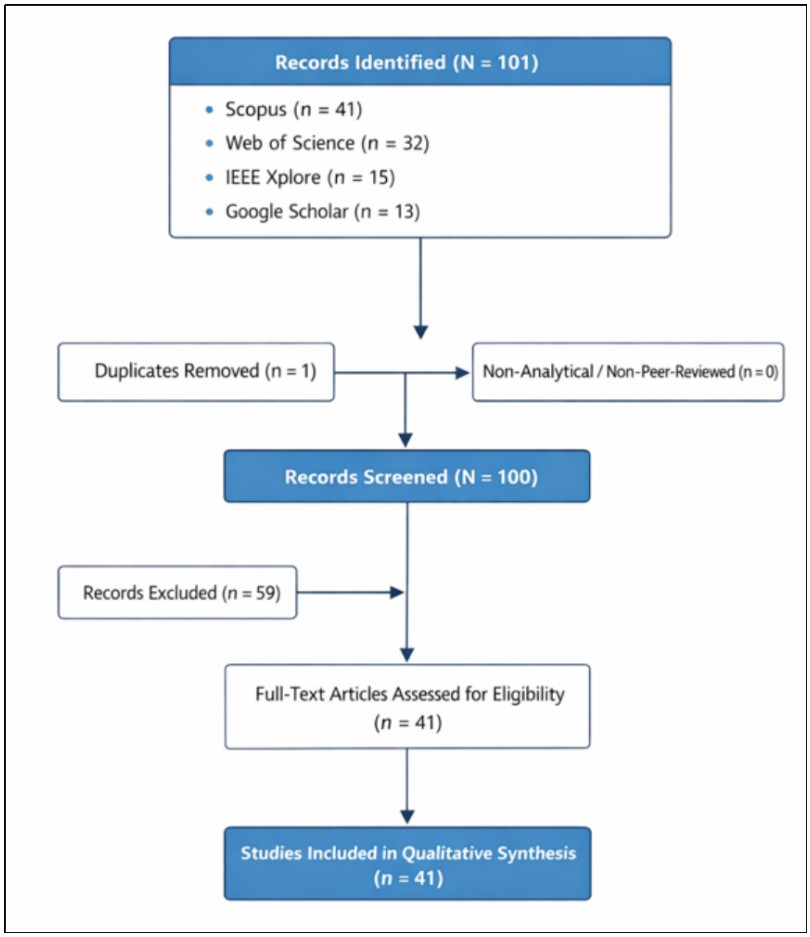

Figure 1. PRISMA flow diagram illustrating the study selection process.

4.3. Data extraction and analysis

Data extraction was conducted using a structured template capturing study objectives, methodological approach, domain of application, detection techniques, and key findings. To strengthen analytical depth, studies were additionally coded for the type of verification method employed (e.g., pixel-level analysis, metadata forensics, machine learning classifiers, watermarking, blockchain-based approaches) and the context of application (e.g., social media, legal, artistic, forensic, humanitarian). Thematic analysis proceeded across three analytical dimensions:

1) technical verification methods and reported effectiveness,
2) social and behavioural implications, including human perception and misuse patterns,
3) governance, ethical, and regulatory considerations.

Where reported in the literature, quantitative indicators such as detection accuracy ranges, frequency of method adoption, or comparative performance claims were noted to support a limited quantitative synthesis. While heterogeneity in

study designs precluded formal meta-analysis, this approach enabled comparative evaluation of strengths, limitations, and applicability contexts of different detection strategies, supporting a more analytically grounded synthesis rather than purely descriptive aggregation. To enhance evidence evaluation, included studies were also subjected to a qualitative quality appraisal, considering criteria such as methodological transparency, clarity of evaluation metrics, dataset robustness, and relevance to real-world deployment. This appraisal informed the weighting of evidence in the synthesis, ensuring that conclusions reflect both the breadth and relative strength of the existing literature.

4.4. Illustrative case analysis

Four purposively selected case studies were examined to contextualise the literature findings within real-world settings, specifically social media platforms where AI-generated images are actively created, shared, and moderated. These cases were chosen for their analytical richness rather than representativeness, providing detailed insight into the interplay of technical verification tools, user behaviour, platform governance, and institutional responses.

Each case was mapped to key components of the conceptual framework, illustrating the application of technical verification methods, the influence of social and behavioural dynamics, and the ethical and governance challenges associated with AI-generated visual content. The analyses reveal patterns that are difficult to capture through literature alone, such as perceptual biases, deliberate misuse of synthetic visuals, and operational constraints faced by platforms and regulators.

Together, the thematic literature synthesis and the illustrative case analyses establish a rigorous, multi-layered methodological approach. This approach treats AI image verification not merely as a technical problem, but as a socio-technical challenge shaped by complex interactions across technology, society, and institutional governance.

5. Thematic analysis and input to the conceptual models

5.1. Political disinformation

AI-generated synthetic visuals have emerged as potent tools in political disinformation, enabling the fabrication of images and videos that manipulate voter perceptions and influence electoral outcomes. Deepfakes have been deployed to misattribute statements, stage events, or undermine political figures, highlighting vulnerabilities in democratic processes and public trust [19]. Despite advances in forensic detection, including pixel-level analysis and machine learning classifiers, verification remains reactive rather than preventive, often identifying manipulated content only after widespread circulation [24]. This underscores the need for integrative strategies combining AI-powered detection, platform-level moderation, and civic digital literacy to anticipate and mitigate political misinformation. From a conceptual standpoint, political disinformation illustrates the intersection of technological capability, societal impact, and governance, reinforcing the socio-technical framing of

AI image verification.

5.2. Corporate and financial deception

In corporate contexts, synthetic visuals facilitate sophisticated financial fraud through executive impersonation, forged contracts, or falsified evidence, threatening investor confidence and organisational integrity [20]. Forensic frameworks leveraging machine learning, metadata analysis, and blockchain-enabled verification have demonstrated efficacy in detecting such manipulations, yet they require continual adaptation to evolving generative techniques [27, 28]. The literature suggests that mitigating these risks also depends on aligning technological tools with regulatory oversight and internal governance structures, highlighting a critical socio-technical interdependency. Corporate deception conceptualises how AI image verification must bridge technical detection and strategic risk management within high-stakes, dynamic environments.

5.3. Social media manipulation

Social media platforms amplify the reach and velocity of synthetic visuals, creating fertile ground for misinformation, identity fraud, and viral content distortion [21]. Deepfake images and videos can rapidly influence perceptions, public debates, and trending narratives, often outpacing the capacity of human moderators. Hybrid detection methods combining AI-driven classifiers, watermarking, and blockchain verification offer promising resilience, but scalability and user accessibility remain challenges [29]. Conceptually, social media manipulation demonstrates the co-evolution of user behaviour, platform governance, and technological interventions, reinforcing the study's socio-technical perspective and highlighting the need for integrated, adaptive verification systems.

5.4. Academic and scientific misinformation

AI-generated visuals and synthetic data present emerging risks in academic and scientific contexts, where manipulated figures, charts, or experimental outputs can distort findings and mislead peer review [11, 12]. Such manipulations threaten the integrity of knowledge production and compromise trust in scholarly communication. Detection methods, including forensic image analysis and provenance tracking, are essential but insufficient in isolation; embedding verification protocols into research workflows and institutional policies is critical. This theme underscores the broader relevance of AI image verification beyond public or commercial domains, situating authenticity as a cross-sectoral challenge requiring systemic safeguards.

5.5. Legal and forensic evidence tampering

Synthetic visuals increasingly pose challenges in legal and forensic contexts, where falsified images or video evidence can undermine judicial processes and investigative accuracy [25]. Techniques such as Discrete Fourier Transform-based SVM classifiers and blockchain-based watermarking provide robust tools for detecting tampered evidence. Still, their operational integration requires coordination between

technological developers, legal authorities, and forensic practitioners [26]. Conceptually, this theme highlights the critical interplay between AI detection technologies, legal standards, and institutional capacity, emphasising the socio-technical complexity of safeguarding evidentiary integrity.

5.6. Humanitarian and crisis misinformation

In humanitarian and crisis contexts, AI-generated visuals can exacerbate panic, spread false aid claims, or distort situational awareness, directly affecting response effectiveness and public perception [8]. Rapid verification is crucial to prevent misallocation of resources or reputational harm to agencies. Emerging solutions include real-time AI detection, cross-platform coordination, and digital literacy interventions to empower responders and the public. This thematic focus illustrates the societal stakes of synthetic visual verification. It reinforces the argument that effective AI image verification strategies must integrate technical, ethical, and organisational considerations to protect vulnerable populations.

6. Case study analysis

Recent case studies highlight the complexities and innovations in validating and checking AI-generated images against human-made ones. Velásquez-Salamanca, Martín-Pascual, and Andreu-Sánchez [38] found that individuals frequently misclassify AI-generated visuals as human-made, with human-created content rated as more realistic and credible. This underscores the difficulty of authenticity judgments in digital contexts. On social media, Ricker et al. [39] conducted a large-scale analysis of 15 million X (formerly Twitter) profiles, revealing that although only 0.052% used AI-generated images, such accounts often engaged in spamming and political amplification, demonstrating the strategic use of synthetic visuals for manipulation. In the creative domain, Salas Espasa and Camacho argued that AI-generated art challenges long-held concepts of authenticity, leading to a hybrid "aura" that combines human intention with machine execution [40]. However, debates persist about its emotional and artistic depth. Complementing these perspectives, Ediboglu Bartos and Özmen Akyol showed that deep learning methods increasingly distinguish between synthetic and real images, offering a promising technological avenue for addressing risks associated with generative visuals [41]. These studies illustrate both the challenges of human perception and the potential of machine learning in safeguarding authenticity across social, political, and artistic contexts.

The case studies summarised in **Table 2** highlight the complexity of validating AI-generated images against human-made ones. While individuals frequently misclassify AI-generated visuals and assign greater credibility to authentic images [38], large-scale analyses reveal their increasing use in online spaces, particularly for manipulation and disinformation [39]. At the same time, AI reshapes artistic authenticity [40], and deep learning models show strong potential for effective detection [41].

Table 2. Simplified case study analysis.

Study	Focus	Key insight
Velásquez-Salamanca et al. [38]	Human interpretation of images	People often mistake AI images for real ones; real images are more credible.
Ricker et al. [39]	X (formerly Twitter) profile images	A small but significant share of profiles use AI-generated faces, often for spamming or manipulation.
Salas Espasa and Camacho [40]	AI in art	AI challenges traditional ideas of authenticity, creating a hybrid "aura" of human–machine creativity.
Ediboglu Bartos and Özmen Akyol [41]	Deep learning authentication	AI models can reliably detect fake images better than human perception

7. Conceptual framework

The conceptual framework developed in this study (**Figure 2**) illustrates the interplay between AI-generated images, human perception, and verification mechanisms. Case study evidence indicates that individuals frequently misclassify AI-generated visuals as human-made, whereas technical approaches such as deep learning classifiers, metadata analysis, and digital watermarking offer effective detection strategies. The performance of these verification mechanisms is influenced by socio-technical factors, including digital literacy, professional expertise, platform policies, regulatory oversight, and ethical norms, which can enhance or constrain their effectiveness. The framework emphasizes that achieving trust and authenticity in digital media requires not only robust detection tools but also coordinated engagement across technological, organizational, and societal domains, highlighting its integrative rather than purely novel contribution relative to existing socio-technical and responsible-AI frameworks.

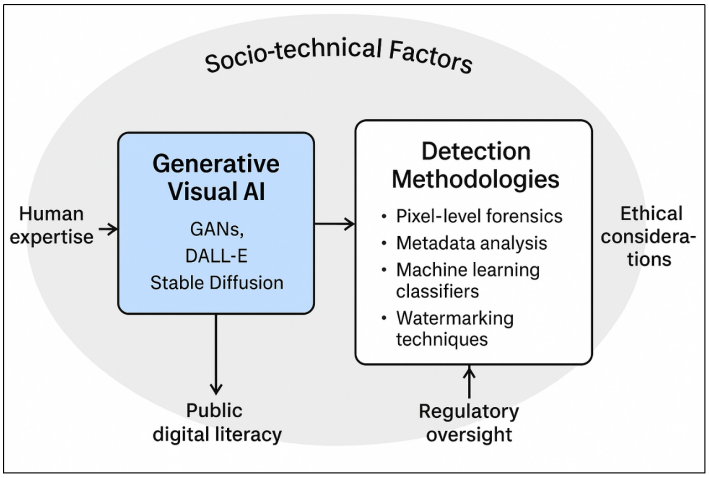

Figure 2. Framework for AI-generated image detection and verification.
Source: Own construct.

8. Key findings and discussion

This study examined the challenges and opportunities associated with validating AI-generated images compared to human-made visuals, with a focus on social media,

artistic, and forensic contexts. The findings address the broader aim of understanding how individuals and technological systems perceive, classify, and authenticate synthetic visuals, highlighting both technical and socio-technical dimensions. As generative AI tools increase in sophistication, the ability to accurately detect and interpret their outputs is critical for safeguarding trust, originality, and integrity across multiple domains.

Case study evidence reveals a consistent perceptual bias: individuals frequently misclassify AI-generated images as human-made, while authentic human-created visuals are often rated as more realistic and credible [38]. This underscores the limitations of unaided human judgment in establishing authenticity and raises important questions about how trust is formed in visual media. Large-scale analyses in social media contexts further show that even a small proportion of AI-generated content, for example, the 0.052% of AI-generated profile images detected on X (formerly Twitter), can be exploited for spamming, political amplification, and disinformation campaigns, demonstrating disproportionate societal risks relative to prevalence [39].

Beyond issues of deception, AI-generated visuals are reshaping traditional concepts of creativity and authenticity in the arts. AI-mediated creation challenges Walter Benjamin's notion of the "aura," resulting in hybrid forms of authenticity that merge human intent with machine execution [40]. This evolution offers opportunities to democratize creative expression while simultaneously raising concerns over perceived emotional depth and originality, highlighting that verification is not only a technical challenge but also a cultural and ontological one.

In response to these challenges, computational detection methods provide promising solutions. Deep learning models, metadata analysis, and hybrid verification approaches have demonstrated superior accuracy in distinguishing AI-generated from authentic images, outperforming human perception [41]. These results indicate that while human judgment remains fallible, technical verification tools can supplement or replace human evaluation in scenarios demanding precision and speed.

The conceptual framework developed in this study (**Figure 2**) integrates these insights by mapping the interactions between AI-generated images, human perception, and verification mechanisms. It emphasizes the influence of socio-technical factors such as digital literacy, professional expertise, platform governance, regulatory oversight, and ethical norms on the effectiveness of detection mechanisms. This framework highlights that ensuring authenticity and trust in digital media requires not only robust technological solutions but also coordinated engagement across organizational, societal, and regulatory domains.

Novelty and Practical Implications: Unlike prior socio-technical or responsible-AI frameworks, this study offers an integrative perspective that explicitly connects human perceptual biases, technical detection methods, and governance mechanisms. Its practical implications are manifold: social media platforms, policymakers, and content creators can leverage the framework to design targeted verification strategies, enhance digital literacy, and develop ethical guidelines, thereby mitigating misinformation, protecting reputations, and reinforcing public trust in digital imagery.

9. Study limitations

While this study comprehensively analyses AI image verification across social media, artistic, and forensic contexts, several limitations warrant consideration. The reliance on secondary data from literature and case studies constrains the ability to observe real-time behavioural interactions with AI-generated visuals. Additionally, the focus on selected platforms and technologies may not fully capture the diversity of generative AI applications or the evolving sophistication of deepfakes. Methodologically, the integration of thematic literature review and qualitative case study approaches, while robust, cannot establish causal relationships, and contextual and demographic variations in public perceptions, platform governance, and technological adoption limit the generalizability of findings.

10. Recommendations

This study recommends combining technological innovation, regulatory oversight, and digital literacy initiatives to strengthen authenticity and trust in digital media. Technical measures should include ongoing development and deployment of hybrid detection systems, leveraging deep learning, metadata analysis, watermarking, and blockchain verification. Simultaneously, organisations and governments should implement clear governance frameworks, ethical standards, and platform-level policies to mitigate manipulation and disinformation risks. Public education campaigns are essential to enhance digital literacy, empowering users to assess visual content critically. At the same time, cross-disciplinary collaboration between technologists, policymakers, and social scientists can ensure adaptive, context-sensitive approaches to AI image verification.

11. Conclusions

This study demonstrates that AI-generated images present transformative opportunities and significant challenges across social, political, corporate, academic, and humanitarian domains. Human perception alone cannot discern authenticity reliably, making technical verification methods essential for safeguarding trust, privacy, and integrity. The conceptual framework developed highlights the interdependence of technological, organisational, and socio-cultural factors in adequate image verification. The findings underscore that addressing the risks of synthetic visuals requires coordinated, socio-technical strategies that integrate algorithmic innovation, governance structures, and user engagement to maintain credibility and resilience in digital media ecosystems.

Funding: This research and manuscript preparation were conducted without external funding or financial support.

Institutional review board statement: This study did not involve human participants or animal subjects and thus did not require ethical approval.

Informed consent statement: Not applicable.

Data availability statement: This study exclusively used data obtained from secondary sources through a comprehensive literature review. All referenced data are publicly accessible and have been appropriately cited within the manuscript.

Conflict of interest: The author declares no conflict of interest.

AI use statement: While preparing this manuscript, the author used Grammarly and QuillBot, solely for English language editing and polishing. All suggestions provided by these tools were carefully reviewed and revised by the author, who take full responsibility for the accuracy, clarity, and integrity of the final content presented in this publication.

References

1. Xu Z. Innovative applications of artificial intelligence in new media content creation and dissemination. Frontiers in Humanities and Social Sciences. 2025; 5(2): 85–91. doi: 10.54691/dz2hgr20
2. Zhao Y. The synergistic effect of artificial intelligence technology in the evolution of visual communication of new media art. Heliyon. 2024; 10(18): e38008. doi: 10.1016/j.heliyon.2024.e38008
3. Chi J. The evolutionary impact of artificial intelligence on contemporary artistic practices. Communications in Humanities Research. 2024; 35(1): 6–11. doi: 10.54254/2753-7064/35/20240006
4. Swargiary K, Roy K. Exploring the impact of artificial intelligence on visual arts: Technological advancements, market dynamics, ethical considerations, and human creativity. 2024. Available online: https://www.researchgate.net/publication/381530797_EXPLORING_THE_IMPACT_OF_ARTIFICIAL_INTELLIGENCE_ON_VISUAL_ARTS_TECHNOLOGICAL_ADVANCEMENTS_MARKET_DYNAMICS_ETHICAL_CONSIDERATIONS_AND_HUMAN_CREATIVITY
5. Sordo Z, Chagnon E, Hu Z, et al. Synthetic scientific image generation with VAE, GAN, and diffusion model architectures. Journal of Imaging. 2025; 11(8): 252. doi: 10.3390/jimaging11080252
6. Mahajan S, Gite S, Pradhan B, et al. Integrating speech-to-text for image generation using generative adversarial networks. CMES—Computer Modeling in Engineering and Sciences. 2025; 143(2): 2001–2026. doi: 10.32604/cmes.2025.058456
7. Zhang Y, Pang Z, Huang S, et al. Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. Journal of King Saud University—Computer and Information Sciences. 2025; 37: 148. doi: 10.1007/s44443-025-00154-8
8. Bengesi S, El-Sayed H, Sarker MK, et al. Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Access. 2024; 12: 69812–69837. doi: 10.1109/ACCESS.2024.3397775
9. Hemraj S. AI and the future of creative development: Redefining digital media production. AI and Ethics. 2025; 5: 5105–5119. doi: 10.1007/s43681-025-00765-x
10. Al-kfairy M, Mustafa D, Kshetri N, et al. Ethical challenges and solutions of generative AI: An interdisciplinary perspective. Informatics. 2024; 11(3): 58. doi: 10.3390/informatics11030058
11. Sohail S, Sajjad SM, Zafar A, et al. Deepfake image forensics for privacy protection and authenticity using deep learning. Information. 2025; 16(4): 270. doi: 10.3390/info16040270
12. Astillero R. Forensic analysis of image metadata to distinguish AI-generated images. Zenodo. 2025. doi: 10.5281/zenodo.16871919
13. Solanke AA, Biasiotti MA. Digital forensics AI: Evaluating, standardising and optimising digital evidence mining techniques. KI—Künstliche Intelligenz. 2022; 36: 143–161. doi: 10.1007/s13218-022-00763-9
14. Balasubramaniam S, Chirchi V, Kadry S, et al. The road ahead: Emerging trends, unresolved issues, and concluding remarks in generative AI: A comprehensive review. International Journal of Intelligent Systems. 2024; 2024: 4013195. doi: 10.1155/2024/4013195
15. Radanliev P, Santos O, Ani UD. Generative AI cybersecurity and resilience. Frontiers in Artificial Intelligence. 2025; 8: 1568360. doi: 10.3389/frai.2025.1568360

16. Thampanichwat C, Wongvorachan T, Sirisakdi L, et al. Mindful architecture from text-to-image AI perspectives: A case study of DALL-E, Midjourney, and Stable Diffusion. Buildings. 2025; 15(6): 972. doi: 10.3390/buildings15060972
17. Yıldırım E. Comparative analysis of Leonardo AI, Midjourney, and DALL-E: AI's perspective on future cities. Urbanizm. 2023; 28: 82–96.
18. Babaei R, Cheng S, Duan R, et al. Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. Journal of Sensor and Actuator Networks. 2025; 14(1): 17. doi: 10.3390/jsan14010017
19. Hynek N, Gavurova B, Kubak M. Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries. Journal of Innovation and Knowledge. 2025; 10(5): 100782. doi: 10.1016/j.jik.2025.100782
20. Kaushik P, Garg V, Priya A, et al. Financial fraud and manipulation: The malicious use of deepfakes in business. In: Deepfakes and Their Impact on Business. IGI Global; 2024. pp. 173–196. doi: 10.4018/979-8-3693-6890-9.ch008
21. Sharma A, Pujari M, Goel A. The rise of AI-generated synthetic identities: A new frontier in social media. International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences. 2025; 13(2). doi: 10.37082/IJIRMPS.v13.i2.232397
22. Vecchietti G, Liyanaarachchi G, Viglia G. Managing deepfakes with artificial intelligence: Introducing the business privacy calculus. Journal of Business Research. 2025; 186: 115010. doi: 10.1016/j.jbusres.2024.115010
23. Romero-Moreno F. Deepfake detection in generative AI: A legal framework proposal to protect human rights. Computer Law and Security Review. 2025; 58: 106162. doi: 10.1016/j.clsr.2025.106162
24. Ghiurău D, Popescu DE. Distinguishing reality from AI: Approaches for detecting synthetic content. Computers. 2025; 14(1): 1. doi: 10.3390/computers14010001
25. Ölvecký M, Host'ovecky M. Digital image forensics using EXIF data of digital evidence. In: Proceedings of the 19th International Conference on Emerging eLearning Technologies and Applications (ICETA); 11–12 November 2021; Košice, Slovakia. doi: 10.1109/ICETA54173.2021.9726649
26. Ferreira S, Antunes M, Correia ME. Exposing manipulated photos and videos in digital forensics analysis. Journal of Imaging. 2021; 7(7): 102. doi: 10.3390/jimaging7070102
27. Darwish SM, Abu-Deif MM, Elkaffas SM. Blockchain for video watermarking: An enhanced copyright protection approach for video forensics based on perceptual hash function. PLoS One. 2024; 19(10): e0308451. doi: 10.1371/journal.pone.0308451
28. Mastoi QUA, Memon MF, Jan S, et al. Enhancing deepfake content detection through blockchain technology. International Journal of Advanced Computer Science and Applications. 2025; 16(6): 48–58. doi: 10.14569/IJACSA.2025.0160607
29. Suryawanshi V, Desai TJ, Shinde K, et al. The art of detection: Methods for identifying AI-generated visual content. International Journal of Creative Research Thoughts. 2025; 13(6). doi: 10.56975/ijcrt.v13i6.288562
30. Amerini I, Barni M, Battiato S, et al. Deepfake media forensics: Status and future challenges. Journal of Imaging. 2025; 11(3): 73. doi: 10.3390/jimaging11030073
31. Aziz B, Blackwell C, Islam S. A framework for digital forensics and investigations: The goal-driven approach. International Journal of Digital Crime and Forensics. 2013; 5(2): 1–22. doi: 10.4018/jdcf.2013040101
32. Bokolo B, Liu Q. Artificial intelligence in social media forensics: A comprehensive survey and analysis. Electronics. 2024; 13(9): 1671. doi: 10.3390/electronics13091671
33. Lombardo G, Mordonini M, Tomaiuolo M. Adoption of social media in socio-technical systems: A survey. Information. 2021; 12(3): 132. doi: 10.3390/info12030132
34. Tangi L, Rodriguez Müller AP, Janssen M. AI-augmented government transformation: Organisational transformation and the socio-technical implications of artificial intelligence in public administrations. Government Information Quarterly. 2025; 42(3): 102055. doi: 10.1016/j.giq.2025.102055
35. Al-Debei MM, Avison D. Business model requirements and challenges in the mobile telecommunication sector. Journal of Organisational Transformation and Social Change. 2011; 8(2): 215–235. doi: 10.1386/jots.8.2.215_1
36. Savaget P, Geissdoerfer M, Kharrazi A, et al. The theoretical foundations of socio-technical systems change for sustainability: A systematic literature review. Journal of Cleaner Production. 2019; 206: 878–892. doi: 10.1016/j.jclepro.2018.09.208
37. Priya A. Case study methodology of qualitative research: Key attributes and navigating the conundrums in its application. Sociological Bulletin. 2020; 70(1): 003802292097031. doi: 10.1177/0038022920970318
38. Velásquez-Salamanca D, Martín-Pascual MÁ, Andreu-Sánchez C. Interpretation of AI-generated vs. human-made

images. Journal of Imaging. 2025; 11(7): 227. doi: 10.3390/jimaging11070227
39. Ricker J, Assenmacher D, Holz T, et al. AI-generated faces in the real world: A large-scale case study of Twitter profile images. In: Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses; 30 September–2 October 2024; New York, NY, USA. doi: 10.1145/3678890.3678922
40. Salas Espasa D, Camacho M. From aura to semi-aura: Reframing authenticity in AI-generated art: A systematic literature review. AI and Society. 2025; 40: 6727–6759. doi: 10.1007/s00146-025-02361-3
41. Ediboglu Bartos G, Özmen Akyol S. Deep learning for image authentication: A comparative study on real and AI-generated image classification. In: Proceedings of the 18th International Symposium on Applied Informatics and Related Areas; November 2023; Székesfehérvár, Hungary.